Robust adaptive estimation of dimension reduction space

Pavel Čížek

Wolfgang Härdle

Center for Applied Statistics and Econometrics, Humboldt-Universität zu Berlin, Spandauer Str. 1, 10178 Berlin, Germany.

Summary.
Most dimension reduction methods based on nonparametric smoothing are highly sensitive to outliers and to data coming from heavy tailed distributions. We show that the recently proposed MAVE and OPG methods by Xia et al. (2002) allow us to make them robust in a relatively straightforward way that preserves all advantages of the original approach. The best of the proposed robust modifications, which we refer to as MAVE-WMAD-R, is sufficiently robust to outliers and data from heavy tailed distributions, it is easy to implement, and surprisingly, it also outperforms the original method in small sample behaviour even when applied to normally distributed data.

Keywords: nonparametric regression, dimension reduction, minimum average variance estimator, robust estimation, median absolute deviation, L_1 regression

1. Introduction

In regression, we aim to estimate the regression function, which is the expectation of a dependent variable $y \in \mathbb{R}$ conditional on explanatory variables $X \in \mathbb{R}^p$. This expectation, $E(y|X = x)$, can be, without prior knowledge, modelled nonparametrically. An increasing number of explanatory variables makes nonparametric estimation suffer from the curse of dimensionality. There are two main approaches to deal with high dimensional X variables: we can either assume a simpler form of the regression function, for example its additivity, or we can try to reduce the dimension of the space of explanatory variables. The latter, more general approach received a lot of attention recently, see Li (1991) and Xia et al. (2002), for instance, and it is also in the focus of our interest here.

A dimension-reduction regression model can be written as

$$y = g(B_0^\top X) + \varepsilon, \tag{1}$$

where g is an unknown smooth link function, B_0 represents a $p \times D$ orthogonal matrix, $D \leq p$ and $E(\varepsilon|X) = 0$ almost surely. For $D = p$, we obtain the standard regression model with all explanatory variables X entering independently. Provided that $D < p$, the regression function depends on X only through D linear combinations of explanatory variables X. Hence, to explain the dependent variable y, the space of p explanatory variables X can be reduced to a space given by B_0 with a smaller dimension D. The vectors of B_0 are called directions in this context. The dimension reduction methods aim to find the dimension D of the reduction space and a matrix B_0 defining this space.

Recently, Xia et al. (2002) proposed a new method, the minimum average variance estimation (MAVE), that overcomes several problems of other existing estimators, such as sliced inverse regression (SIR), Li (1991). First, in contrast to other methods, MAVE does not need undersmoothing when estimating the link function g in order to achieve a faster rate of consistency. Second, MAVE can be applied to many models including time series data. Moreover, Xia et al. (2002) show how their approach can be used to generalise some existing methods; for example, they propose the outer product of gradients (OPG) estimator, which extends the average derivative estimator of Härdle and Stoker (1989) to multi-index models. Finally, Xia et al. (2002)'s experience as they indicated in their discussed paper is that MAVE is also robust against outliers in data.

Although MAVE improves over the existing methods both from its convergence and applicability points of view, we doubt that it might be sufficiently robust to withstand outliers in data. The main reason is that it is based on local polynomial smoothing, that is, on local least-squares estimation, which is highly sensitive to outlying observations. One can naturally argue that since the estimation is done locally the estimator is not sensitive to outlying observations in the space of explanatory variables X. On the other hand, the local character of estimation significantly raises possible effects of outliers in the dependent variable y, because the samples used for local estimation of the regression function are rather small. Similar sensitivity to outliers, although in the space of explanatory variables, was observed in the case of SIR by Gather, Hilker, and Becker (2001), who also proposed its robust version. SIR is sensitive to outliers in the explanatory variables since it uses an inverse

regression. Now, because of the vast range of advantages of MAVE and OPG methods, we would like to examine their main weakness—non-robustness to outlying observations—in more details and to propose ways to improve them without affecting their main strengths. From now on, we mean by outliers those observations that are outlying in the dependent variable.

The rest of the paper is organised as follows. In Section 2, we describe both the MAVE and OPG methods and demonstrate their low robustness. Then we propose possible robust enhancements of the methods in Section 3 and compare them by means of simulations in Section 4.

2. Estimation of dimension reduction space

In this section, we present the MAVE and OPG methods as well as a procedure for determining the effective dimension reduction by means of cross-validation. At the end of the section, we will motivate our concerns about robustness of these methods by a small simulation.

2.1. The MAVE method

Let d represent now the working dimension, $1 \leq d \leq p$, where p denotes the number of explanatory variables X. For an assumed number d of directions in model (1) and known directions B_0, one would typically minimise

$$\min E\{y - E(y|B_0^\top X)\}^2$$

to obtain a nonparametric estimate of the regression function $E(y|B_0^\top X)$. The MAVE method is based on the local linear regression, which hinges at a point X_0 on linear approximation

$$E(y|B_0^\top X) \approx a + b^\top B_0^\top (X - X_0). \tag{2}$$

Now, if directions B_0 are not known, we have to search their approximation B. Xia et al. (2002) propose to plug-in unknown directions B in the local linear approximation of the regression function and to optimise simultaneously with respect to B and local parameters a and b of local linear smoothing. Hence, given a sample $(X_i, y_i)_{i=1}^n$ from (X, y), they perform

local linear regression at every $X_0 = X_i, i = 1, \ldots, n$, and end up minimising

$$\min_{\substack{B:B^\top B=I_p \\ a_j,b_j,j=1,\ldots,n}} \sum_{i=1}^{n}\sum_{j=1}^{n}[y_i - \{a_j + b_j^\top B^\top (X_i - X_j)\}]^2 w_{ij}, \tag{3}$$

where I_p represents the $p \times p$ identity matrix and w_{ij} are weights describing local character of linear approximation (2) (i.e., w_{ij} should depend on the distance of points X_i and X_j).

Xia et al. (2002) call the resulting estimator of B MAVE and show that the simultaneous minimisation with respect to local linear approximation given by a_j, b_j and to directions B results in a convergence rate superior to any other dimension-reduction method. Initially, a natural choice of weights is given by a multidimensional kernel function K_h. At a given X_0,

$$w_{i0} = K_h(X_i - X_0) \Big/ \sum_{i=1}^{n} K_h(X_i - X_0) \tag{4}$$

for $i = 1, \ldots, n$ and a kernel function $K_h(\cdot)$, where h refers to a bandwidth. Additionally, when we already have an initial estimate of the dimension reduction space given by $\hat{B}$, it is possible to iterate and search an improved estimate of the reduction space. Xia et al. (2002) do so by using the initial estimator $\hat{B}$ to measure distances between points X_i and X_0 in the reduced space. More precisely, they propose to choose in the iterative step weights

$$w_{i0} = K_h\{\hat{B}^\top (X_i - X_0)\} \Big/ \sum_{i=1}^{n} K_h\{\hat{B}^\top (X_i - X_0)\}\ . \tag{5}$$

Repeating such iteration steps until convergence results in a refined MAVE (rMAVE) estimator. From now on, whenever we talk or refer to MAVE, we mean its refined version rMAVE.

2.2. The OPG method

Based on the above described MAVE approach, Xia et al. (2002) also manage to generalise the average derivate estimator (ADE) by Härdle and Stoker (1989) to more dimensions. Instead of using the moment condition for the gradient of the regression function g in model (1), $E\{\nabla g(X)\} = 0$, they start from the average outer product of gradients (OPG), $E\{\nabla g(X) \nabla^\top g(X)\}$. By decomposing the MAVE objective function, it can be shown that the searched dimension reduction matrix B consists of the d eigenvectors corresponding to

the d largest eigenvalues of $E\{\nabla g(X)\nabla^\top g(X)\}$. Now, recalling once again that local linear fitting solves for a given sample $(X_i, y_i)_{i=1}^n$ and a given point $X_j, j \in \{1, \ldots, n\}$

$$\min_{a_j, b_j} \sum_{i=1}^{n} [y_i - \{a_j + b_j^\top (X_i - X_j)\}]^2 w_{ij}, \tag{6}$$

we can estimate $\Sigma = E\{\nabla g(X)\nabla^\top g(X)\}$ by

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \hat{b}_j^\top \hat{b}_j,$$

where $\hat{b}_j$ are estimates of b_j from (6). Hence, the OPG method consists in estimating $\hat{\Sigma}$ and determining its d eigenvectors corresponding to the d largest eigenvalues. Also this method can be iteratively refined in the same way as MAVE by determining weights w_{ij} in (6) using the distance of observations in the reduced space once its initial estimate is known. Similarly to MAVE, whenever we talk about the OPG method, we mean its refined version.

The OPG method generalises the average derivate estimation, but it does not reach the rate of consistency of the MAVE method. Apart from being an interesting generalisation, we mention it here because it is easy to implement and to modify as we will see later. Moreover, our initial simulations showed that it can perform as well as MAVE in small samples and in the presence of outliers; see Section 4 for more details.

2.3. Dimension of effective reduction space

The described methods are capable of estimating the dimension reduction space provided we can specify its dimension. To determine the dimension d, Xia et al. (2002) extend the cross-validation approach of Yao and Tong (1994). The cross-validation criterion is defined as

$$CV(d) = \sum_{j=1}^{n} \left[y_j - \sum_{i=1, i\neq j}^{n} \frac{y_i K_h\{\hat{B}^\top (X_i - X_j)\}}{\sum_{i=1, i\neq j}^{n} K_h\{\hat{B}^\top (X_i - X_j)\}} \right]$$

for $d > 0$ and for the special case of independent y and X as

$$CV(0) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

Consequently, the dimension is then determined as

$$\hat{d} = \text{argmin}_{0 \leq d \leq p} CV(d),$$

Table 1. Estimates of the EDR dimension with one outlier. Table (a) reports frequencies of the best EDR dimension found. Table (b) considers frequencies of the best EDR dimensions only if the conditional variance of the dependent variable on the indices is lower than the unconditional variance, otherwise zero is reported.

	Outlier value			
Dimension	200	400	600	800
1	21	19	14	11
2	21	21	31	29
3	34	36	31	30
4	16	12	9	7
5	5	6	3	4
6	2	2	4	2
7	1	1	2	2
8	0	1	2	0
9	0	2	3	9
10	0	0	1	6

(a)

	Outlier value			
Dimension	200	400	600	800
0	21	77	96	96
1	17	2	1	2
2	17	5	2	1
3	22	9	1	2
4	17	3	0	0
5	4	3	0	0
6	2	3	0	0
7	1	1	0	0
8	0	0	0	0
9	0	0	0	0

(b)

where p represents the number of explanatory variables.

Using this cross-validation procedure, let us now motivate by a small simulation our concerns regarding the robustness of the MAVE method. Consider the following nonlinear model

$$y_i = (X_i^\top b_1)^2 - (0.5 + X_i^\top b_2)^2 + 15\cos(X_i^\top b_3) + 0.5\varepsilon_i,$$

where all random variables have the standard normal distribution in $\mathbb{R}^{10}$ and $b_1 = (1, 2, 3, 0, 0, 0, 0, 0, 0, 0)/\sqrt{14}$, $b_2 = (-2, 1, 0, 1, 0, 0, 0, 0, 0, 0)/\sqrt{6}$, and $b_3 = (0, 0, 0, 0, 0, 0, 0, 1, 1, 1)/\sqrt{3}$. Additionally, we include one observation that has the value y_i replaced by a constant between 200 and 800. The frequencies of estimated EDR dimensions for 100 repetitions and sample size $n = 100$ are summarised in Table 1 (a) and (b): the former plainly reports the best dimension found (without considering $CV(0)$), whereas the latter contains either the best dimension found if the corresponding model was able to explain at least a part of the variance of the dependent variable or zero (with the $CV(0)$ definition employed). Apparently, the further the outlier is, the more cases are not correctly identified and estimated (zero entries in Table 1 (b) actually say that the method has not found any dependence

between y and X variables). Moreover, one can notice that the outlier leads generally to an upward biased EDR dimension. Hence, we see that the MAVE method and the cross-validation based on it can be significantly influenced by a single outlying observation.

3. Robust enhacements

In the previous sections, we have argued that the MAVE and OPG methods can be highly sensitive to outliers in data, mainly because the local linear regression is based on simple least squares. Therefore, we would like to propose several possible enhancements of the MAVE and OPG method that should increase their robust properties, optimally without affecting their other qualities. There are two strategies that can be used in this case: first, we can design weights w_{ij} in (3) depending on y_i values to reduce effects of outlying observations; second, we can replace the local least squares fitting by a more robust procedure. In this section, we describe both strategies and their variants, postponing their finite-sample comparison to Section 4.

3.1. Robust choice of weights

The easiest way to make the discussed dimension reduction methods more robust to outliers is certainly an adjustment of weights w_{ij} in (3), since it does not require any principal change of the methods or the corresponding computational procedures. Let us remind that the initial choice of weights (4) proposed by Xia et al. (2002) is given at some $X_0 \in \mathbb{R}^p$ by

$$w_{i0} = K_h(X_i - X_0) \Big/ \sum_{i=1}^{n} K_h(X_i - X_0)$$

for $i = 1, \ldots, n$ and a kernel function $K_h(\cdot)$. Hence, the observations distant in the space of explanatory variables X are downweighted by the kernel function K_h anyway and we have to take care only about outlying observations in the direction of the dependent variable y.

A natural way to determine observations that are outlying and to downweight them is to measure locally (at a point X_0) the mean $\hat{\mu}_y$ and standard deviation $\hat{\sigma}_y$ of y-values. Then, for given values y_i we can decrease weights of observations indirectly proportional to the normalised values $t_i = |y_i - \hat{\mu}_y|/\hat{\sigma}_y$. Optionally, we can set weights equal to zero for observations with $t_i > K$, where $K > 0$ is a suitable constant (for example, $K = 3$), to avoid their influence completely. Now, although the arithmetical mean and standard

deviation are standard measures of location and scale, their sensitivity to outliers hints that they do not have to present a very reasonable choice. Thus, we employ their more robust equivalents—the median and median absolute deviation (MAD)—as well; see (Hampel et al., 1986, pp. 105, 106, 235) explaining why MAD is a suitable robust estimator of scale when scale is a nuisance parameter. Summing up these ideas, we obtain four possible choices of initial weights w_{i0} for an observation $(X_i, y_i), i = 1, \ldots, n$, and a point X_0.

Weighted standard deviation without rejection (WSTD)

Let us define the weighted mean $\hat{\mu}_y(X_0)$ at X_0

$$\hat{\mu}_y(X_0) = \sum_{i=1}^{n} \frac{y_i K_h(X_i - X_0)}{\sum_{i=1}^{n} K_h(X_i - X_0)}$$

and the weighted standard deviation

$$\hat{\sigma}_y(X_0) = \sqrt{\sum_{i=1}^{n} \frac{\{y_i - \hat{\mu}_y(X_0)\}^2 K_h(X_i - X_0)}{\sum_{i=1}^{n} K_h(X_i - X_0)}}.$$

Then set weights to

$$w_{i0} = \frac{K_h(X_i - X_0)}{\sum_{i=1}^{n} K_h(X_i - X_0)} \cdot \frac{\hat{\sigma}_y(X_0)}{\max\{|y_i - \hat{\mu}_y(X_0)|, \hat{\sigma}_y(X_0)\}}.$$

Weighted standard deviation with rejection (WSTD-R)

Using the previously defined weighted mean $\hat{\mu}_y(X_0)$ and weighted standard deviation $\hat{\sigma}_y(X_0)$, set

$$w_{i0} = \frac{K_h(X_i - X_0)}{\sum_{i=1}^{n} K_h(X_i - X_0)} \cdot \frac{\hat{\sigma}_y(X_0)}{\max\{|y_i - \hat{\mu}_y(X_0)|, \hat{\sigma}_y(X_0)\}} \cdot I\{|y_i - \hat{\mu}_y(X_0)| \leq 3\hat{\sigma}_y(X_0)\}.$$

Weighted median absolute deviation without rejection (WMAD)

Let us define the weighted median $\tilde{\mu}_y(X_0)$ at X_0

$$\tilde{\mu}_y(X_0) = \min_{k=1,\ldots,n} \left\{ y_{(k)} \left| \sum_{i=1}^{n} \frac{K_h(X_i - X_0)}{\sum_{i=1}^{n} K_h(X_i - X_0)} \cdot I\left(y_i \leq y_{(k)}\right) \geq 0.5 \right. \right\},$$

where $y_{(k)}$ represents the kth order statistics of the sample $\{y_i\}_{i=1}^{n}$ and $[\cdot]$ denotes the integer part. Analogously, define the weighted median absolute deviation

$$\tilde{\sigma}_y(X_0) = 1.4826 \cdot \min_{k=1,\ldots,n} \left\{ r_{(k)} \left| \sum_{i=1}^{n} \frac{K_h(X_i - X_0)}{\sum_{i=1}^{n} K_h(X_i - X_0)} \cdot I\left(r_i \leq r_{(k)}\right) \geq 0.5 \right. \right\},$$

where $r_i = |y_i - \tilde{\mu}_y(X_0)|$. Then the weights are defined by

$$w_{i0} = \frac{K_h(X_i - X_0)}{\sum_{i=1}^n K_h(X_i - X_0)} \cdot \frac{\tilde{\sigma}_y(X_0)}{\max\{|y_i - \tilde{\mu}_y(X_0)|, \tilde{\sigma}_y(X_0)\}}.$$

Weighted median absolute deviation with rejection (WMAD-R)

Using the previously defined weighted median $\tilde{\mu}_y(X_0)$ and weighted median absolute deviation $\tilde{\sigma}_y(X_0)$, set

$$w_{i0} = \frac{K_h(X_i - X_0)}{\sum_{i=1}^n K_h(X_i - X_0)} \cdot \frac{\tilde{\sigma}_y(X_0)}{\max\{|y_i - \tilde{\mu}_y(X_0)|, \tilde{\sigma}_y(X_0)\}} \cdot I\{|y_i - \tilde{\mu}_y(X_0)| \leq 3\tilde{\sigma}_y(X_0)\}.$$

Similarly to the original MAVE and OPG methods, the robust weights can also be interactively refined. Having an initial estimate $\hat{B}$ of the dimension reduction space, we can measure the distances between points X_i and X_0 in the reduced space. Analogously to (5), we can then define the refined weights, for example, for the WMAD weights as follows: the weighted mean at X_0

$$\hat{\mu}_y(X_0|\hat{B}) = \sum_{i=1}^n \frac{y_i K_h\{\hat{B}^\top(X_i - X_0)\}}{\sum_{i=1}^n K_h\{\hat{B}^\top(X_i - X_0)\}},$$

the weighted standard deviation

$$\hat{\sigma}_y(X_0|\hat{B}) = \sqrt{\sum_{i=1}^n \frac{\{y_i - \hat{\mu}_y(X_0|\hat{B})\}^2 K_h\{\hat{B}^\top(X_i - X_0)\}}{\sum_{i=1}^n K_h\{\hat{B}^\top(X_i - X_0)\}}},$$

and the refined weights

$$w_{i0} = \frac{K_h\{\hat{B}^\top(X_i - X_0)\}}{\sum_{i=1}^n K_h\{\hat{B}^\top(X_i - X_0)\}} \cdot \frac{\hat{\sigma}_y(X_0|\hat{B})}{\max\{|y_i - \hat{\mu}_y(X_0|\hat{B})|, \hat{\sigma}_y(X_0|\hat{B})\}}.$$

3.2. Other robust methods

A further strategy how robust properties of MAVE and OPG can be improved consists in replacing the local least square regression by a more robust method. There are plenty of robust regression methods and some, such as smoothed least trimmed squares by Čížek (2001), would suit MAVE and OPG methods very well. The only, but important limitation is the speed of computation of such robust methods, which significantly limits their applicability in this case (it is necessary to solve at least kdn regression problems, where typically $k > 10$). Nevertheless, since we perform regression only locally, it suffices to use a method robust only to outlying observations in the direction of the dependent variable y. Hence,

to meet the requirements on speed and robustness, we propose to use local L_1 regression instead of local least squares; see (Hampel et al., 1986, Secs. 6.2, 6.4).

Consequently, in the case of MAVE, we try to estimate local L_1 regression for a sample $(X_i, y_i)_{i=1}^n$ and $X_0 = X_i, i = 1, \ldots, n$ by minimising

$$\min_{\substack{B:B^\top B=I_p \\ a_j,b_j,j=1,\ldots,n}} \sum_{i=1}^{n}\sum_{j=1}^{n} \left|y_i - \{a_j + b_j^\top B^\top (X_i - X_j)\}\right| w_{ij}. \tag{7}$$

We refer further to this method as MAVE-L1. Similarly, in the case of OPG, we estimate local L_1 regression at all points $X_j, j \in \{1, \ldots, n\}$,

$$\min_{a_j,b_j} \sum_{i=1}^{n} \left|y_i - \{a_j + b_j^\top (X_i - X_j)\}\right| w_{ij}, \tag{8}$$

and we compute $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \hat{b}_j^\top \hat{b}_j$. It remains to determine the d eigenvectors of $\hat{\Sigma}$ corresponding to the d largest eigenvalues. This method is further referred to as OPG-L1.

Note that whereas the OPG-L1 can be easily implemented (it is just local L_1 regression), the computation of MAVE-L1 presents serious difficulties. The iterative process proposed by Xia et al. (2002) for the original MAVE method relies on alternating minimisation with respect to (a_j, b_j) and B. Whereas the first case, minimisation with respect to (a_j, b_j) for a given directions B, is nothing but local L_1 regression, the minimisation problem for $B = (\beta_1, \ldots, \beta_d)$ given $(a_j, b_j) = (a_j, b_{j1}, \ldots, b_{jd})$ has to be rewritten in the following way:

$$\begin{aligned}
& \min_{B:B^\top B=I_p} \sum_{i=1}^{n}\sum_{j=1}^{n} \left|y_i - \{a_j + b_j^\top B^\top (X_i - X_j)\}\right| w_{ij} \\
= & \min_{B:B^\top B=I_p} \sum_{i=1}^{n}\sum_{j=1}^{n} \left| y_i - \left\{ a_j + \sum_{k=1}^{d} b_{jk}\beta_k^\top (X_i - X_j) \right\} \right| w_{ij} \\
= & \min_{B:B^\top B=I_p} \sum_{i=1}^{n}\sum_{j=1}^{n} \left| y_i - \left\{ a_j + \sum_{k=1}^{d} \beta_k^\top b_{jk} (X_i - X_j) \right\} \right| w_{ij}.
\end{aligned}$$

This represents a regression problem with n^2 observations and pd variables, and thus, its size will be enormous as the sample size increases. On the other hand, there are very fast algorithms available for computing L_1 regression in large data sets, see for example Koenker and Portnoy (1997). We use here the implementation of L_1 estimation in the statistical environment XploRe.

The proposed approach for computing MAVE-L1 can be also used for computing the original MAVE method. Since our proposal differs from that of Xia et al. (2002), we will compare MAVE computed in both ways. To differentiate, the simulations using the algorithm of Xia et al. (2002) are referred to plainly by MAVE, whereas the simulations using the algorithm proposed in this section are labelled MAVE-ALT.

4. Simulations

In this section, we will compare the original MAVE and OPG method with their modifications proposed in Section 3 by means of simulations. First, we introduce the models used for simulations. Next, we explain why we actually use and compare both MAVE and OPG here in spite of the fact that results of Xia et al. (2002) show MAVE being superior to OPG. Finally, we compare the original OPG and MAVE methods and their proposed modifications using simulations.

4.1. Simulation models

Throughout this section, we consider the following nonlinear model (used already in Section 2.3)

$$y_i = (X_i^\top \beta_1)^2 - (0.5 + X_i^\top \beta_2)^2 + 15\cos(X_i^\top \beta_3) + 0.5\varepsilon_i, \tag{9}$$

where all explanatory variables have the standard normal distribution in $\mathbb{R}^{10}$ and $\beta_1 = (1, 2, 3, 0, 0, 0, 0, 0, 0, 0)/\sqrt{14}$, $\beta_2 = (-2, 1, 0, 1, 0, 0, 0, 0, 0, 0)/\sqrt{6}$, and $\beta_3 = (0, 0, 0, 0, 0, 0, 0, 1, 1, 1)/\sqrt{3}$. To compare the robust properties of all estimators, we use three variants of this model.

(a) The standard normal case $\varepsilon_i \sim N(0, 1)$ serves for a comparison of methods when no outlying observations are present. Moreover, it corresponds to one of the simulation settings used by Xia et al. (2002).

(b) The Student distributed errors $\varepsilon_i \sim t_1$ with one degree of freedom simulate case where there is a higher probability of larger errors, but the (heavier-tailed) error distribution is symmetric and centred around zero.

(c) The standard normal errors $\varepsilon_i \sim N(0, 1)$ are used for 95% of observations, while the remaining 5% of observations are outliers with y-values generated from the uniform

distribution $U(-600, 600)$. This combination simulates a normal data contaminated by several large outliers that are not related with the original model at all.

For the sake of brevity, we refer further to these three cases as NORMAL, STUDENT, and OUTLIERS, respectively.

For all models in all simulations, we use sample size $n = 100$ and 100 repetitions (we observed that the results for larger samples sizes, such as $n = 200$, are qualitatively the same as for $n = 100$). Moreover, all variants of MAVE and OPG applied to these models were created by modifying existing MAVE and OPG algorithms available in statistical environment XploRe. The methods use the Gaussian kernel by default.

Finally, let us note that to compare the methods, we use the same distance measure of the estimated space $\hat{B}$ and the true space $B_0 = (\beta_1, \beta_2, \beta_3)$ as Xia et al. (2002): $m(\hat{B}, B_0) = \|(I - B_0 B_0^T)\hat{B}\|$ for $d \leq D = 3$ and $m(\hat{B}, B_0) = \|(I - \hat{B}\hat{B}^T)B_0\|$ for $d \geq D = 3$ and ($D = 3$ is the true dimension of the reduced space used in our simulations, whereas d denotes the dimension used for estimation).

4.2. MAVE vs. OPG

Let us now explain why we consider both the MAVE and derived OPG methods. The main reason is that it is hard to argue theoretically which method will be more stable and robust under various circumstances. For example, using the three models introduced in Section 4.1 we simulate 100 data sets, and assuming the correct dimension $d = 3$, estimate them by both the (refined) MAVE and OPG methods. The average estimation errors $m(\hat{B}, B_0)$ for various bandwidth choices decomposed to $m(\hat{\beta}_1, B_0)$, $m(\hat{\beta}_2, B_0)$, and $m(\hat{\beta}_3, B_0)$ are depicted in Figure 1, whereby OPG and MAVE are represented by solid and dashed lines, respectively. Although the MAVE method is certainly preferable for clean data (case NORMAL), in correspondence with the results of Xia et al. (2002), OPG seems to perform better in the case of model OUTLIERS, although both OPG and MAVE fits are rather poor in this case. It is hard to judge in the case of the STUDENT model. Consequently, we cannot a priori decide, which method suits some data better. Moreover, although MAVE has a higher convergence rate, OPG offers easy implementation and fast computation.

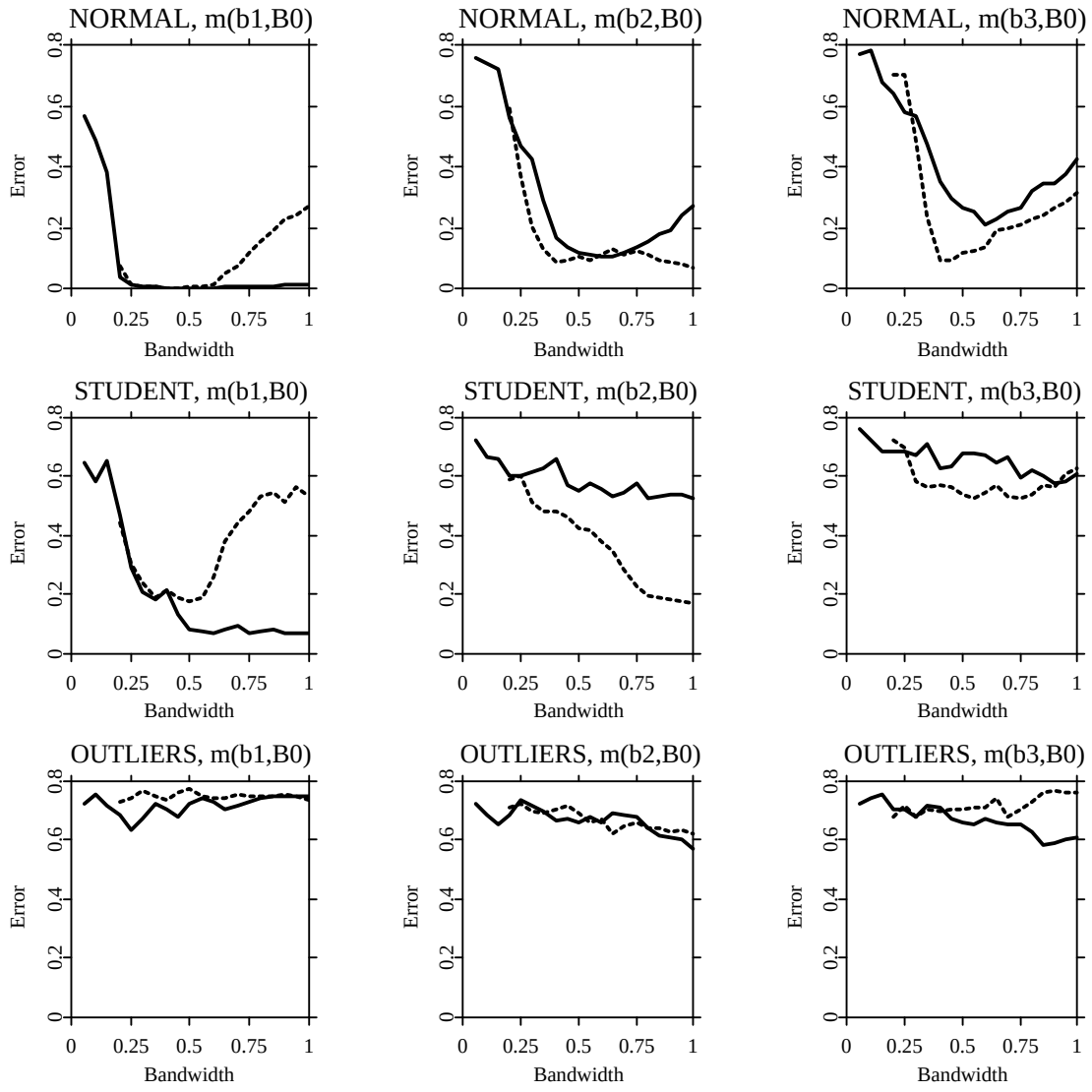

Fig. 1. Average errors of MAVE and OPG for models NORMAL, STUDENT and OUTLIERS and all parameters β_1, β_2, and β_3 of model (9). The solid line represents OPG, the dashed line MAVE.

4.3. OPG simulations

Now, let us compare the original OPG method with all its proposed modifications. We use again models NORMAL, STUDENT, and OUTLIERS. We generate for every model

Table 2. Median and mean average errors of the OPG estimates of the dimension reduction space. Mean average errors are in brackets.

Data	Method	Parameter		
		$m(\hat{\beta}_1, B_0)$	$m(\hat{\beta}_2, B_0)$	$m(\hat{\beta}_3, B_0)$
NORMAL	OPG original	0.0043 (0.0051)	0.1057 (0.1623)	0.2135 (0.3344)
	OPG WSTD	0.0034 (0.0045)	0.1135 (0.1687)	0.2014 (0.3013)
	OPG WSTD-R	0.0037 (0.0046)	0.1197 (0.1722)	0.2524 (0.3634)
	OPG WMAD	0.0038 (0.0045)	0.0905 (0.1470)	0.1885 (0.3395)
	OPG WMAD-R	0.0039 (0.0044)	0.1063 (0.1523)	0.3033 (0.3999)
	OPG L1	0.0054 (0.0065)	0.1333 (0.1924)	0.2389 (0.3175)
STUDENT	OPG original	0.0691 (0.2714)	0.5595 (0.5343)	0.6696 (0.6516)
	OPG WSTD	0.0600 (0.2332)	0.5593 (0.5250)	0.6063 (0.5928)
	OPG WSTD-R	0.0348 (0.1973)	0.5008 (0.5100)	0.6473 (0.5883)
	OPG WMAD	0.0451 (0.2018)	0.6180 (0.5599)	0.6211 (0.5700)
	OPG WMAD-R	0.0266 (0.1455)	0.5381 (0.5273)	0.6735 (0.6266)
	OPG L1	0.0163 (0.0994)	0.4547 (0.4465)	0.4957 (0.5546)
OUTLIERS	OPG original	0.7297 (0.7041)	0.6609 (0.6404)	0.6697 (0.6522)
	OPG WSTD	0.7105 (0.6798)	0.4873 (0.5029)	0.5329 (0.5479)
	OPG WSTD-R	0.5906 (0.5942)	0.4571 (0.4605)	0.5663 (0.5569)
	OPG WMAD	0.0153 (0.1006)	0.4617 (0.4693)	0.5110 (0.5230)
	OPG WMAD-R	0.0135 (0.1299)	0.4287 (0.4344)	0.6102 (0.5846)
	OPG L1	0.0073 (0.0084)	0.1801 (0.2590)	0.3732 (0.4019)

100 samples and estimate them using OPG, OPG-WSTD, OPG-WSTD-R, OPG-WMAD, OPG-WMAD-R, and OPG-L1; see Section 3 for the description of these methods. The median and mean estimation errors $m(\hat{B}, B_0)$ decomposed to $m(\hat{\beta}_1, B_0)$, $m(\hat{\beta}_2, B_0)$, and $m(\hat{\beta}_3, B_0)$ are presented in Table 2.

First, we discuss results in the case of model NORMAL. In this case, methods do not differ too much from each other. Nevertheless, it is interesting to notice that the weighted variants of OPG, OPG-WSTD and OPG-WMAD, are slightly better than the original OPG method. This can be an effect of relatively small samples used in this simulation ($n = 100$). The worst, although the difference is not very high, is the L_1 based OPG-L1 method.

Second, let us look at the simulations for the data generated from model STUDENT,

which are by definition more scattered and may contain larger errors than data coming from model NORMAL. In this case, the original OPG method is apparently the worst one although the simple weighted version OPG-WSTD does not differ much. The modifications employing rejection of too distant observations, OPG-WSTD-R and OPG-WMAD-R, perform better than the non-rejecting variants. The best method is however the OPG-L1 method, which clearly outperforms all other methods.

Third, the situation changes again once we analyse the performance of the methods for the OUTLIERS model, which contain 5% of random noise with a large amplitude. The original OPG method fails for all parameters (the maximum value of the error $m(\hat{\beta}, B_0)$ is one). The modifications downweighting observations using the weighted standard deviation, OPG-WSTD and OPG-WSTD-R, are slightly better, but also unsatisfactory. On the other hand, the methods using robust estimates of location and scale, OPG-WMAD and OPG-WMAD-R, are able to identify the first parameter vector well and are certainly preferable to the original OPG method. Altogether, all these methods are again outperformed by the OPG-L1 method, which is significantly better.

Finally, we can conclude that if slightly worse performance of the OPG-L1 method in the standard normal case does not matter, OPG-L1 provides best results when the data contain larger errors or outlying observations. Otherwise, OPG-WMAD(-R) can be recommended since they are easily implementable, sufficiently robust in all cases, and OPG-WMAD even outperforms the original OPG for normally distributed data in small samples.

4.4. MAVE simulations

Let us now compare the original MAVE method with all its proposed modifications. We use again models NORMAL, STUDENT, and OUTLIERS. We generate for every model 100 samples and estimate them using MAVE, MAVE-WSTD, MAVE-WSTD-R, MAVE-WMAD, MAVE-WMAD-R, MAVE-ALT, and MAVE-L1; see Section 3 for the description of these methods. The median and mean estimation errors $m(\hat{B}, B_0)$ decomposed to $m(\hat{\beta}_1, B_0)$, $m(\hat{\beta}_2, B_0)$, and $m(\hat{\beta}_3, B_0)$ are presented in Table 3.

As we can see, the results are qualitatively similar to those for OPG. Most importantly, the original MAVE method is outperformed by its modifications in all cases NORMAL, STUDENT, and OUTLIERS. MAVE-WMAD and MAVE-WMAD-R can be considered the

Table 3. Median and mean average errors of the MAVE estimates of the dimension reduction space. Mean average errors are in brackets.

Data	Method	Parameter		
		$m(\hat{\beta}_1, B_0)$	$m(\hat{\beta}_2, B_0)$	$m(\hat{\beta}_3, B_0)$
NORMAL	MAVE original	0.0051 (0.0128)	0.1038 (0.1587)	0.1205 (0.2204)
	MAVE WSTD	0.0042 (0.0084)	0.0734 (0.1203)	0.0999 (0.1777)
	MAVE WSTD-R	0.0042 (0.0083)	0.0903 (0.1497)	0.0953 (0.2191)
	MAVE WMAD	0.0032 (0.0069)	0.0608 (0.1315)	0.0818 (0.1928)
	MAVE WMAD-R	0.0031 (0.0121)	0.0752 (0.1301)	0.0804 (0.1876)
	MAVE ALT	0.0055 (0.0305)	0.0536 (0.1088)	0.0736 (0.1585)
	MAVE L1	0.0059 (0.0212)	0.0888 (0.1722)	0.1499 (0.3050)
STUDENT	MAVE original	0.1790 (0.2984)	0.4245 (0.4416)	0.5415 (0.5273)
	MAVE WSTD	0.0707 (0.2049)	0.4277 (0.4620)	0.5606 (0.5449)
	MAVE WSTD-R	0.0778 (0.2464)	0.4385 (0.4485)	0.5332 (0.5333)
	MAVE WMAD	0.0622 (0.1597)	0.4989 (0.4973)	0.5155 (0.5268)
	MAVE WMAD-R	0.0989 (0.1921)	0.4360 (0.4571)	0.4265 (0.4546)
	MAVE ALT	0.4151 (0.4053)	0.4286 (0.4411)	0.5364 (0.5122)
	MAVE L1	0.0944 (0.1872)	0.2805 (0.3551)	0.3977 (0.4625)
OUTLIERS	MAVE original	0.7424 (0.7075)	0.6688 (0.6405)	0.7124 (0.6801)
	MAVE WSTD	0.7002 (0.6707)	0.4856 (0.4931)	0.6904 (0.6709)
	MAVE WSTD-R	0.6397 (0.6233)	0.4474 (0.4723)	0.7072 (0.6688)
	MAVE WMAD	0.0487 (0.1159)	0.3616 (0.3691)	0.3505 (0.3879)
	MAVE WMAD-R	0.0476 (0.1153)	0.2350 (0.3154)	0.3621 (0.4004)
	MAVE ALT	0.6858 (0.6805)	0.7187 (0.6859)	0.7479 (0.6847)
	MAVE L1	0.0595 (0.1814)	0.1863 (0.2677)	0.2267 (0.3243)

best ones from methods using the original algorithm by Xia et al. (2002) since they perform well in all three cases. One has to realize, however, that the sample size is $n = 100$. We expect the difference between MAVE and its modifications to disappear with an increasing sample size for the NORMAL model.

It is also interesting to compare MAVE and MAVE-ALT, since MAVE-ALT performs better than MAVE for the NORMAL model (assuming the same number of MAVE refinements in both cases). On the other hand, MAVE-ALT is slightly worse when applied to contaminated data (OUTLIERS and STUDENT cases). A debate which algorithm is preferable in practice is very easy to decide—the computational method by Xia et al. (2002) is typically 20–30 times faster than MAVE-ALT.

Finally, let us look at the performance of MAVE-L1. Similarly to OPG, MAVE-L1 does not excel in the case of the NORMAL model, although the difference between MAVE-L1 and other methods is relatively small. On the other hand, it performs better than all the other methods once applied to contaminated data (the STUDENT and OUTLIERS models). In the case of MAVE, its L1 modification is unfortunately disadvantaged by rather slow computation and high memory demands (it uses the same algorithm as MAVE-ALT).

Consequently, we can conclude that, the difference between MAVE-WMAD-R and MAVE-L1 being significant, but not extremely large, the MAVE-WMAD-R is probably the best recommendation for everyday use.

4.5. MAVE vs. OPG revised

Let us now compare the MAVE and derived OPG methods in the same way as in Subsection 4.2, but this time considering their best modifications. Thus, we compare OPG-WMAD-R and OPG-L1 with MAVE-WMAD-R (MAVE-L1 is excluded because of its impractically high computational demands). Using the three models NORMAL, STUDENT, and OUTLIERS introduced in Section 4.1 we simulate 100 data sets, and assuming the correct dimension $d = 3$, estimate them by all three methods. The average estimation errors $m(\hat{B}, B_0)$ for various bandwidth choices decomposed to $m(\hat{\beta}_1, B_0)$, $m(\hat{\beta}_2, B_0)$, and $m(\hat{\beta}_3, B_0)$ are depicted in Figure 2, whereby OPG-WMAD-R, OPG-L1, and MAVE-WMAD-R are represented by thin solid, thick dashed, and thin dashed lines, respectively.

First, a general observation is that the OPG method is slightly better in determining

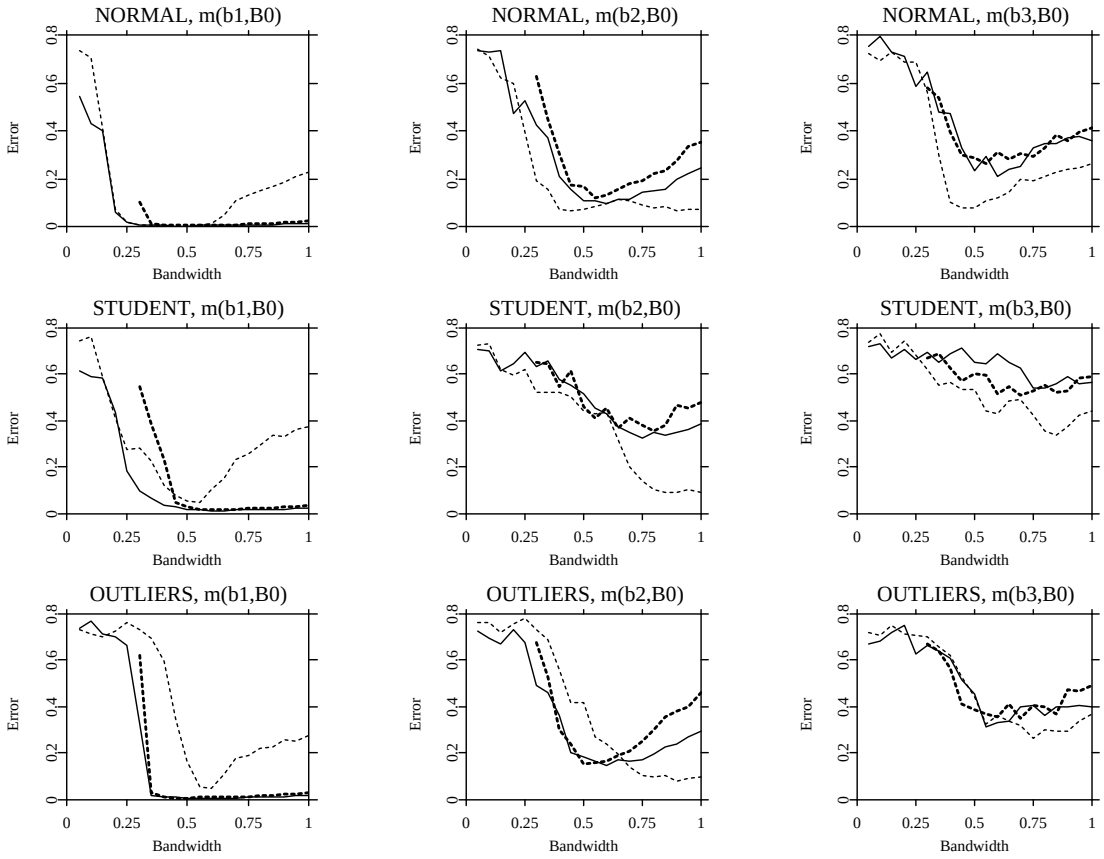

Fig. 2. Average errors of MAVE-WMAD-R, OPG-WMAD-R, and OPG-L1 for models NORMAL, STUDENT and OUTLIERS and all parameters β_1, β_2, and β_3 of model (9). The thin solid line represents OPG-WMAD-R, the thin dashed line MAVE-WMAD-R, and the thick dashed line represents OPG-L1.

the first direction, see the graphs of $m(\hat{\beta}_1, B_0)$ in Figure 2, whereas the MAVE method estimates the remaining directions with (often substantially) smaller errors, see for example the graphs of $m(\hat{\beta}_2, B_0)$) for the STUDENT model and of $m(\hat{\beta}_3, B_0)$ for the NORMAL model in Figure 2. This observation is consistent with results of Xia et al. (2002). Overall performance of MAVE measured by $m(\hat{B}, B_0)$ is typically better than that of OPG.

Second, MAVE-WMAD-R is clearly preferable for clean data (case NORMAL) and it outperforms the modifications of OPG in all cases when we take into account estimation

errors for all directions. Thus, MAVE-WMAD-R proves to be sufficiently robust both to outliers and to data from heavy tailed distributions. Combined with the superior performance of MAVE-WMAD-R for clean data (it even outperforms the original MAVE method in small samples), MAVE-WMAD-R is the best modification of MAVE proposed here. Additionally, the fact that it can be computed in the same way as the original MAVE method by Xia et al. (2002), and is thus easy to implement, makes it attractive for practical use.

Finally, let us note that also OPG methods might be a good choice if a fast computation is highly desirable, since computing OPG is usually several times faster than the equivalent MAVE method for the same data. Moreover, OPG-L1 might be preferred when data are supposed to come from an exponential-type distribution.

5. Conclusion

In this paper, we address the robustness properties of dimension reduction methods. Most dimension reduction methods that are based on some kind of nonparametric smoothing are highly sensitive to outliers and to data coming from heavy tailed distributions. Although it is in general non-trivial to make dimension reduction methods more robust, we show that the recently proposed MAVE and OPG methods by Xia et al. (2002) allow us to make them robust in a relatively straightforward way that preserves all advantages of Xia et al. (2002)'s approach. Theoretically, the MAVE-L1 modification might be most appealing, especially because of its robustness, but it is handicapped by a very slow computation. Therefore, from the practitioners' point of view, we find that MAVE-WMAD-R is the best of the proposed MAVE and OPG modifications: it is sufficiently robust to outliers and data from heavy tailed distributions, it is easy to implement, and surprisingly, it even outperforms the original MAVE method in small sample behaviour for normally distributed data.

References

Čížek, P. (2001) Robust estimation with discrete explanatory variables. CERGE-EI Working Paper 187/2001.

Gather, U., Hilker, T., and Becker, C. (2001) A Robustified Version of Sliced Inverse

Regression. In Fernholz et al., eds., *Statistics in Genetics and in the Environmental Sciences*, Birkhäusel, Basel, 147–157.

Härdle, W., Hlávka, Z. and Klinke, S. (2000) *XploRe Application Guide*. Heidelberg: Springer-Verlag.

Härdle, W. and Stoker, T. M. (1989) Investigating smooth multiple regression by method of average derivatives. *Journal of American Statistical Association*, **84**, 986–995.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., Stahel, W. A. (1989) *Robust Statistics, The Approach Based on Influence Functions*. United States: John Wiley & Sons.

Horowitz, J. and Härdle, W. (1996) Direct Semiparametric Estimation of a Single-Index Model with Discrete Covariates. *Journal of the American Statistical Association*, **91**, 1632–1640.

Ichimura, H. (1993) Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models. *Journal of Econometrics*, **1993**, 71–120.

Koenker, R., and Portnoy, S. (1997) The Gaussian Hare and the Laplacian Tortoise: Computability of Squared-error vs. Absolute-error Estimators. *Statistical Science*, **12**, 279–300.

Li, K. C. (1991) Sliced inverse regression for dimension reduction. *Journal of American Statistical Association*, **86**, 316–342.

Xia, Y., Tong, H., Li, W. K., and Zhu, L.-X. (2002) An adaptive estimation of dimension reduction space. *Journal of Royal Statistical Society, Series B*, **64 / 3**, 363–410.

Yao, Q., and Tong, H. (1994) On subset selection of in nonparametric stochastic regression. *Statist. Sin.*, **4**, 51–70.